

SOMR — Self-Organizing Map Recommender

Joel Bennett

Golisano College of Computing and Information Sciences
Rochester Institute of Technology

Defense of Master's Project, December 2008

Outline

- 1 Introduction
 - Recommender Systems
 - A Better Recommender
- 2 Background
 - Social Tagging
 - SOM Algorithm
 - PowerShell
- 3 Design
 - Overview
 - Gathering Data
 - SOM Networks
 - Recommender
- 4 Project Outcome
 - Recommendations
 - Shortcomings

Recommender Systems

- Information Filtering
- Content-based Systems
- Collaborative Systems
- Hybrid Systems

A Better Recommender

Research Foundations

- Recommender Systems
A special case of Information Filtering.
- Social Tagging
Users describing items.
- Self-Organizing Maps
User friendly clustering.

Social Tagging

- Public bookmarking
- Tags summarize content
- Free-form keywords

SOM Algorithm

- Unsupervised Clustering
- Vector Space Classifier
- Outstanding Visualization
- Browsable Results

PowerShell

Rapid Algorithm Development

- Command-Line Interface
- Classes as Apps
- Scripts and Pipelines
- Simplest Output

Overview

Data Flow

- Gathering Data
- Generating 2D Maps
- Getting Recent Items
- Rate and Recommend Items

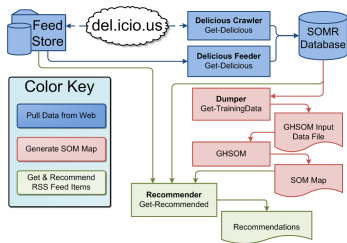
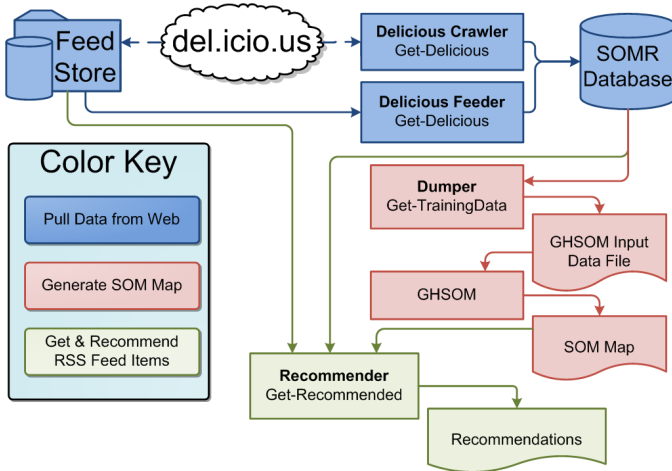


Figure: Data Flow

Overview

Data Flow



Gathering Data

- Fetcher (SgmlParser)
 - Http Scraper
 - Multi-threaded
- Delicious Crawler
 - XPath Parsing
 - Dumps data to database
- SOMR Database
 - Users, Tags, Urls
 - Relationships (Tag-Url-User)

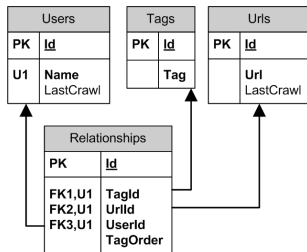
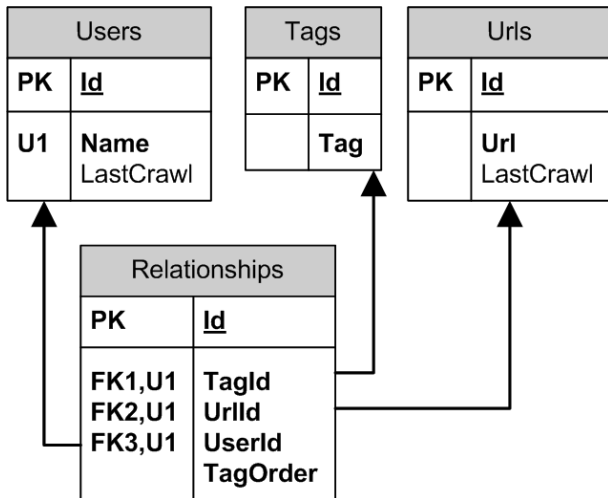


Figure: Database Design

Gathering Data



GHSOM

Accommodating Third-party Code

- Get-TrainingData
- Two Training Sets
- GHSOM executable
 - Takes hours and hours to run

Recommender

- Pre-gather Recent Links
- Scrape URL data
- Determine Users' Interests
- Map URL to Users and Tags

Recommender

- Pre-gather Recent Links
- Scrape URL data
- Determine Users' Interests
- Map URL to Users and Tags
 - On User Map
 - On Tag Map

Recommender

- Pre-gather Recent Links
- Scrape URL data
- Determine Users' Interests
- Map URL to Users and Tags
 - On User Map
 - On Tag Map
 - Find Areas of Interest
 - K-Means Clustering of tags

Recommendations

DEMO: Get-Recommended

- Shows all items, unsorted
- Get-Recommended function
- In my testing...

Recommendations

Sample Output

Rating	Tag R#	User R#	Link
74.793	92.798	49.587	http://www.yammer.com/
66.250	86.335	38.131	http://www.os2world.com/
63.534	86.720	31.072	http://idlebackup.nl/
63.454	81.780	37.796	http://www.ovguide.com/
62.709	77.226	42.385	http://www.osdata.com/
62.246	86.720	27.981	http://umbrellatoday.com/
61.412	81.220	33.680	http://www.stormpulse.com/
61.174	83.896	29.363	http://savefile.com/
60.904	62.440	58.754	http://www.slate.com/id/2195892/
60.449	83.896	27.622	http://www.sendspace.com/
60.009	77.455	35.584	http://www.truecrypt.org/docs/
59.890	74.033	40.090	http://www.kidsastronomy.com/
59.582	81.220	29.289	http://www.americanrhetoric.com/top100sp
59.472	78.394	32.980	http://www.openwinforms.com/
59.245	86.720	20.779	http://www.sweetcron.com/
57.231	81.220	23.646	http://www.techbargains.com/
58.342	83.896	22.567	http://en.wikipedia.org/wiki/Antikythera
57.597	83.896	20.779	http://www.rememberthemilk.com/
57.558	81.780	23.646	http://www.collectivex.com/

Downside

Shortcomings of the current system

- Performance
 - 12 Hours to train GHSOM
 - Throttled Scraping
- User Interface
 - Get-Delicious should be a service
 - Need Graphical UI with clickable links
 - Need Summary Viewer for maps

Upside

Future Work, Opportunities for improvement

- Visualization
- User-Selectable Interest
- Improvements to SOM
- Altered Weighting System

Deliverables

- Source and Binaries for SOMR
- Databases and bookmarks
- Documentation